ASSOCIATION RULE MINING IN E-COMMERCE: A SURVEY

Venkateswari S*

Department of Software Engineering, Noorul Islam University, K.K.Dist, India

Suresh R.M

Department of computer science RMD Engineering College, Kavaraipettai, Chennai, India

Abstract

Association Rule mining is one of the most popular data mining techniques which can be defined as extracting the interesting correlation and relation among large volume of transactions. E-commerce applications generate huge amount of operational and behavioral data. Applying association rule mining in e-commerce application can unearth the hidden knowledge from these data. In this paper a survey of association rule mining and its various applications in e-commerce environment are made.

KEYWORDS

Association Rule mining, frequent itemsets, e-commerce

1. INTRODUCTION

One of the popular descriptive data mining techniques is Association rule mining [1], owing to its extensive use in marketing and retail communities and many other diverse fields. Mining association rules are particularly useful for discovering relationships among items from large databases [2]. The "market-basket analysis" which performs a study on the habits of customers [3] is the source of motivation behind Association rule mining. The extraction of interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories is the main objective of Association rule mining [4].

With the rapid growth of World Wide Web and electronic commerce, huge volume of data is available on line. The available on line data is plenty and with rich descriptions. Since data are electronically collected they are reliable. So e-commerce is a killer domain [17] for Association rule mining. In the e-commerce environment all the actions of customers visiting a shop from entry to exit are recorded. So customer navigation patterns and their purchasing behavior are available in the e-commerce data. Finding association rules from these data helps to make right business decisions in right time. It also helps to improve cross selling and web site design. Also Association rule mining in e-commerce data provide navigation and purchasing suggestions to customers.

This survey is organized as follows. Discussion of Association rule mining and survey of some frequent itemset mining algorithms are made in section 2. Various applications of association rule mining in e-commerce are discussed in section 3. E-commerce data collection is discussed in 4 and conclusion in 5.

2. LITERATURE SURVEY OF FREQUENT ITEMSET GENERATION ALGORITHMS

Association rule mining [3] extracts interesting correlation and relation between large volumes of transactions. This process is divided into two phases. First phase is frequent itemset generation, finding all itemsets that sufficiently exceed minimum support. Second phase is rules construction. From the frequent itemsets generated all association rules having confidence higher than minimum confidence are formed. Frequent itemset generation is the resource consuming task and is the active area of research.

The task of frequent itemset mining was first introduced by Agrawal et al. [3] in 1993. A frequent itemset is a set of items that appears at least in a pre-specified number of transactions. Frequent itemsets are typically used to generate association rules. Apriori [3] was the first and foremost algorithm and forms the foundation for most known algorithms. It states that for a k-itemset to be frequent all its k-1 itemsets have to be frequent. In case of

extremely large input sets, Apriori suffers from two problems of repeated I/O scan and high computational cost. Agrawal et al. [5, 6] proposed the AprioriTid and AprioriHybrid algorithms as well. Park et al. proposed an optimization, called DHP (Direct Hashing and Pruning) intended towards restricting the number of candidate itemsets, shortly following the Apriori algorithms mentioned above [7]. Brin et al put forth the DIC algorithm that segregates the database into intervals of a fixed size so as to reduce the number of traversals through the database [8]. The Eclat algorithm by Zaki [9] is considered to be the archetype in the depth first manner of generation of frequent itemsets. FP-growth algorithm by Han et al. [10] is the most famous and widely used. It is a depth first algorithm. C. Hidber [11] has presented a novel algorithm named CARMA (Continuous Association Rule Mining Algorithm), which is used to compute large itemsets online. A new class of interesting problem called weighted association rule (WAR) was identified by Wei Wang, et al. [12]. They have proposed an approach which mines Weighted Association Rules by first ignoring the weight and finding the frequent itemsets and it was followed by introducing the weight during the rule generation. A new single-pass algorithm, called DSM-FI (Data Stream Mining for Frequent Itemsets) was proposed by Hua-Fu Li, et al. [13], which mines all frequent itemsets over the entire history of data streams. Yu-Chiang Li et al. [14] have evaluated the significance of itemsets for the mining of association rules from databases. They have proposed an algorithm, Enhanced FSM (EFSM), which efficiently reduces the time complexity of the join step. CTU-PRO algorithm was proposed by Alva Erwin et al. [15] to mine the complete set of high utility itemsets from both sparse and relatively dense datasets with short or long high utility patterns. Chun-Jung Chu et al. [16] have proposed a novel method, namely THUI (Temporal High Utility Itemsets)-Mine, used for mining temporal high utility itemsets from data streams efficiently and effectively.

3. APPLICATIONS OF ASSOCIATION RULE MINING IN E-COMMERCE

In this section we survey the articles that implemented association rule mining in e-commerce.

3.1 Web Personalization

Personalization is the use of customer information for delivering a customized solution to that customer thus satisfying personal needs [18, 19]. In the e-commerce environment the available choices for the visiting customers are more. The search cost and time increase due to this overload. Personalization can aid the customer in decision making process. Personalization can also communicate appropriate messages to the right customers on the basis of customer profiles. Typical stages in personalization are discussed by Murthi & sarkar[18]. Common way to incorporate personalization in firm's interaction with customers is through the use of recommender system [20]. How association rule mining is performed on web data to make a web personalized system is discussed by Mobasher[21].

3.2 Recommender systems

Customers like to have a recommender system by which customer can see the feedback from other users who already purchased the products. E-commerce makes use of recommender system to not only show feed back from other users but also suggest interesting and useful products to customers. Geyer et al [22] describes a recommender system that uses association rules derived from past purchases, for making recommendations to new anonymous customers. Diverse recommendation systems are proposed for different business which guides the customers to find products they would like to purchase [23][24]. Most of them are based on either content filtering or collaborative filtering. Content based filtering (CBF) approach recommends products to target customers according to the preferences of their neighbors. The collaborative filtering (CF) approach recommends products to object customers based on their past preferences. The draw backs in these traditional approaches are rectified and an personalized system was proposed by Yiyang zhang e al[25]. Zhang Xizheng[26] propose a personalized recommendation system using association rule mining and classification. Set of association rules are mined from customer requirements database using Apriori algorithm and then he apply CBA-CB algorithm to produce best rules out of whole set of rules.

3.3 Cross selling analysis

The association rule mining is a powerful tool to realize cross selling. Cross selling is a marketing strategy to sell a new product or service to the customer who already used the products of the same enterprise. To introduce a new product or service to a new customer and an old customer, the old customer is more likely to accept it and the success rate is higher. Cross selling is a marketing method which can improve customer value, for the more the relations between the enterprise and the customer, the more dependent the customer will be on the enterprise, and the higher the loyalty will be. Xiao-li Yin [27] discusses how the banks analyze the association relations between the deal activities and other properties like customer age, gender, education, occupation, etc, and can get the result which bank services or financial products the customer will be interested in.

3.4 Purchasing and traveling behavior of customers.

In e-commerce environment finding association rules between purchasing items is very important. There are many algorithms devised in this field [3][7][8][10]. Path traversal pattern mining is the technique that finds most of the navigation behaviors of customers in the e-commerce environment. This information can be used to improve the website design and performance. Navigational suggestions can also be suggested to customers using this information. Many works are carried out in this field [28][29][30]. Yeu-shi-lee et al[31] propose an algorithm IPA that considers both purchasing behavior and travelling patterns of customer at the same time.

4. E-COMMERCE DATA COLLECTION

Various ways are there to collect data relevant to e-commerce [33]. Web server log files, Web server plug-ins (instrumentation), TCP/IP packet sniffing, application server instrumentation are the primary means of collecting data. Most Data mining techniques rely on web server logs. Kohavi et al[17] emphasize the need for data collection at the application server layer and not on the web server in order to support tagging of data meta data that is essential to the discovery process.

Depending on the goals of the analysis, e-commerce data need to be aggregated at different levels of abstraction. Ecommerce data are classified mainly as web usage data, web content data, web structure data and business data [31].

The Web Usage Data. Web usage data is generated by the interactions between the persons browsing a site and the servers on the e-commerce platform. This data can further be divided into log files and business event records. The useful information delivered by web log mainly includes: IP address of the remote host making the request, Date/Time when the request occurs, URI of the object requested, and referrer field. The referrer field consists of important information for marketing purposes, since it can track how people found your site. The information stored in a cookie log helps to ameliorate the connectionless state of web server interactions, enabling servers to track client access across their hosted web pages. A problem is that web logs only contain the name of the page requested, details of the content displayed on the web page may not even be captured by it. In addition, some events also cannot be determined from web logs. (e.g., abandoning shopping carts, searches fail to find any results.) Fortunately, the problems can be solved through collecting data at the application server layer. Application server knows what is on the page, and can log business events.

The Web Content Data. Web content data include HTML/XML pages, web page templates, email templates for campaigns, images, etc.

The Web Structure Data. Web structure data provide the topology of a site, which represent the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. Structure data also include the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page.

The Business Data. Business data include merchandising information like products, assortments, and price list, business rules like promotion rules, and rules for cross-sells and up-sells and sale transaction data.

5. CONCLUSION

In this paper a survey of association rule mining and frequent itemset mining algorithms are presented. How association rule mining can be applied to e-commerce to improve the services of e-commerce enterprises are discussed in detail. The types of e-commerce data and the data collection methods are also discussed.

References

- [1] F. Bodon, 2003. "A Fast Apriori Implementation", In B. Goethals and M. J. Zaki, editors, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Vol.90 of CEUR Workshop Proceedings.
- [2] Yu-Chiang Li, Jieh-Shan Yeh, Chin-Chen Chang, 2005. "Efficient Algorithms for Mining Share-Frequent Itemsets", In Proceedings of the 11th World Congress of Intl. Fuzzy Systems Association.
- [3] R.Agrawal, T.Imielinski, and A.Swami, 1993. "Mining association rules between sets of items in large databases", in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216.
- [4] S. Kotsiantis, D. Kanellopoulos, 2006. "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32, No. 1, pp.71-82.
- [5] Rakesh Agrawal, and Ramakrishnan Srikant, 1994. "Fast Algorithms for Mining Association Rules", In Proceedings of the 20th Int. Conf. Very Large Data Bases, pp. 487-499.
- [6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo, 1996. "Fast discovery of association rules" In U.M. Fayyad, G. Piatetsky Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pp. 307–328. MIT Press.
- [7] J.S. Park, M.-S. Chen, and P.S. Yu, 1995. "An effective hash based algorithm for mining association rules". In Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, volume 24(2) of SIGMOD Record, pp. 175–186. ACM Press.

- [8] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur, 1997. "Dynamic itemset counting and implication rules for market basket data". In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, volume 26(2) of SIGMOD Record, pp. 255-264. ACM Press.
- [9] M.J. Zaki, May/June 2000. "Scalable algorithms for association mining". IEEE Transactions on Knowledge and Data Engineering, 12(3):372-390.
- [10] Han, J., J. Pei, and Y. Yin, 2000. "Mining Frequent Patterns without Candidate Generation" in ACM SIGMOD Int'l Conference on Management of Data.
- [11] C. Hidber, 1999. "Online association rule mining". In A. Delis, C. Faloutsos, and S.Ghandeharizadeh, editors, Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, volume 28(2) of SIGMOD Record, pp. 145–156. ACM Press.
- [12] Wei Wang, jiong yang and philip S. Yu, 2000. "Efficient Mining of Weighted Association Rules (WAR)", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 270 - 274.
- [13] Hua-Fu Li, Suh-Yin Lee and Man-Kwan Shan, 2004. "An Efficient Algorithm for Mining Frequent Itemsets over the Entire History of Data Streams", In Proceedings of the 1st Int'l. Workshop on Knowledge Discovery in Data Streams, pp. 20-24.
- [14] Yu-Chiang Li, Jieh-Shan Yeh, Chin-Chen Chang, 2005. "Efficient Algorithms for Mining Share-Frequent Itemsets", In Proceedings of the 11th World Congress of Intl. Fuzzy Systems Association.
- [15] Alva Erwin, Raj P. Gopalan, N.R. Achuthan, 2007. "A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets", In Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining, Vol. 84, pp. 3-11.
- [16] Chun-Jung Chu, Vincent S. Tseng, Tyne Liang, 2008. "An efficient algorithm for mining temporal high utility itemsets from data streams", Journal of System Software, Vol. 81, No. 7 pp. 1105-1117.
- [17] R. Kohavi, "Mining E-Commerce Data: The Good, the Bad, and the Ugly," In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (KDD'2001) (New York: ACM Press) pp. 8-13.
- [18] Murthi B P S, Sarkar S 2003. "The role of management sciences in research on personalilzation. Manage. Sci 49:1344-1362
 [19] Scafer J B, Konstan J A, Riedl J 2001 "e-commerce recommender applications." Data mining knowledge discovery 5(1/2):115-153. Resnick P, carian H R 1997 recommender systems. Commun. ACM 40(3):56-58
- [20] Mobasher B 2004 "Web usage mining and personalization". In practical handbook of internet computing (ed) M P Singh (CRC Press)
- [21] Geyer-Schulz A, Hashler M 2002 "comparing two recommender algorithms with the help of recommendations by peers". In WEBKDD 2002 - Mining web data for discovering usage patterns and profiles:4th int. workshop(eds) O R zainane J srivastava, M Spiliopoulou, B Masand (revised papers) Lecture notes in computer science LNAI 2703 (Berlin:springer-Verla) PP 137-58.
- [22] Liu Guo-rong, Zhang Xi-zheng, "Collaborative Filtering Based Recommendation system for Product bundling", Proceeding of ICMSE'06, Lille, France, pp.251-254, 2006.
- [23] Chen, H. C., Chen, A. L. P. "A music recommendation system based on music data grouping and user interests", Proceedings of the ACM /CIKM, Atlanta, USA, pp. 231-238,2004.
- [24] Yiyang Zhang, Jianxin (Roger) Jiao, "An associative classification-based recommendation system for personalization in B2C ecommerce applications", Expert Systems with Applications, vol. 33 ,pp. 357-367, 2007.
- [25] Zhan Xizheng "Building personalized recommendation system in E-commerce using Association rule-based mining and classification". Proceedings of 8th ACIS international conference on software engineering, Artificial intelligence, Networking and parallel/Distributed computing, 2007.
- [26] Xiao-li Yin and Xu-sheng Fang," Data Mining Technology Application in Bank CRM", Economic Research Guide, 2009, pp. 112-113.
- [27] M. S. Chen, J. S. Park and P. S. Yu. "Efficient Data Mining for Path Traversal Patterns in a Web Environment." IEEE Transaction on Knowledge and Data Engineering, Vol. 10, No. 2, pp. 209-221, 1998.
- [28] S. J. Yen. "An Efficient Approach for Analyzing User Behaviors in a Web-Based Training Environment. "International Journal of Distance Education Technologies, Vol. 1, No. 4, pp.55-71, 2003.
- [29] M. S. Chen, X. M. Huang and I. Y. Lin. "Capturing User Access Patterns in the Web for DataMining." Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, pp. 345-348, 1999.
- [30] Yue-shi Lee, Show-Jane Yen, Min-chi Hsieh, hi-Hua Tu," Mining web transaction patterns in Electronic commerce Environment" Proceedings of the IEEE international conference on E-commerce Technology for Dynamic E-Business, 2004
- [31] Hong Yu, Xiaolei Huang, Xiaorong Hu, Changxuan Wan, "Knowledge management in E-commerce: A data mining perspective", proceedings of international conference on Management of e-commerce and e-government, 2009.
- [32] N R Srininasa Raghavan, "Data mining in e-commerce: A survey", sadhana vol.30, parts 2&3, April/june 2005, pp 275-289.